Volume No 03, Special Issue No. 01, March 2015 ISSN (online): 2348 – 7550

EFFECTIVE SECURE MINING OF ASSOCIATION RULE WITH SUBGROUP DISCOVERY IN HORIZONTALLY DISTRIBUTED DATABASES

[1]K. Sindhika, [2]P. Ganesh Kumar,

[1]PG Scholar, [2]Assistant Professor,
Department of Information Technology,
Anna University Regional Centre, Coimbatore (India)

ABSTRACT

Data mining can extract important knowledge from large data collections – but sometimes these collections are split among various parties. In Privacy concerns leakage of information can be tackled by using cryptographic techniques namely AES. The methods incorporate cryptographic techniques to minimize the information shared, it may reduce the performance of mining task. Subgroup discovery concept has been enforced to bring top k pattern for faster mining.It finds patterns in the coupling of the databases, without discovering the local databases.

Indexterms : Secure Mining, Subgroup Discovery, Frequent Itemsets, Association Rule.

I INTRODUCTION

In Data mining, association rule is a popular and well researched method for discovering interesting relationsbetween variables in large databases. Piatetsky-Shapiroillustrates observation made on strong rules discoveredin databases using different measures of interestingness and presented accordingly.Agrawal et al introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets through founded strong rules. For example, the rule Found in the sales data of a supermarket would indicate that if a customer buys coffee powder and milk together, there is possibility of buying sugar.These information can be retrieved for decision making in markets for e.g., product placements. In summation to the above example from market basket analysis association rules are utilized today in many application fields including Web usage mining, intrusion detection and Bioinformatics [5].

The two significant basic measures of association rules are support(s) and confidence (c). Support(s) is defined as the proportion of records that contain X unionY to the overall records in the database. The total for each item is augmented by one, whenever the item is crossed over in different transaction in the database during the course of the scanning.Confidence (c) is determined as the ratio of the number of transactions that contain X union Y to the overall records that contain X. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. If the value of support and confidence are minimum then the association rule is said to be stronger.

The Application ofAssociation rule mining in market basket analysis is

- To observe the point of sales transaction.
- From uses information on what customers buy to provide insights into who they are and why theymake certain purchases.
- To predict the product that occurs together.

Subgroup discovery is a method to identify relations between a dependent variable (target variable) and independent variables. For example, consider the subgroup described by ”smoker=true AND family history=positive” for the target variable coronary heart disease=true. Subgroup discovery does not necessarily focus on finding complete relations; instead partial relations, i.e., (small) subgroups with ”interesting” characteristics can be sufficient. The discovered subgroup patterns must essentially satisfy two conditions.

i. Interpretable for the analyst
ii. It should be interesting with respect to the criteria of the user.

Where the Interestingness is typically defined by a quality function, which can take certain statistical or other user-defined quality criteria into account. Interestingness rely on the characteristics of the user's choice.

II RELATED WORK

Secure Mining of Association Rules in Horizontally Distributed Databases:A novel protocol UNIFI-KC (Unifying lists of locally Frequent Item sets— Kantarcioglu and Clifton) is based on the Fast Distributed Mining (FDM) Algorithm which is an unsecured distributed version of the Apriori algorithm [1]. The main ingredients in our protocol are two novel secure multi-party algorithms - one that computes the union of private subsets that each of the interacting players hold, and other that checks for the inclusion of an element held by one player in a subset held by another. It offers enhanced privacy with respect to the protocol. It is also simpler and significantly more efficient in terms of communication rounds, communication and computational cost.

A Fast Distributed Algorithm for Mining Association Rules: Fast Distributed Mining of association rule which generates a small number of candidate sets and substantially reduces the number of messages to be passed at mining association rules. The FDM algorithm proceeds as follows: (i) Initialization, (ii) Candidate sets generation, (iii) Local Pruning, (iv) Unifying the candidate item sets, (v) Computing local supports and (vi) Broadcast Mining Results. Main discussion made on two issues first one about the relationship between the effective FDM and distribution of data and another one deals with the support threshold relaxation for possible reduction of message overheads. Some interesting properties between global and local large itemsets are observed. Two powerful pruning techniques are proposed namely global pruning and local pruning which could improve the computation speed at data mining preprocessing task [3].

Privacy Preserving Association Rule Mining in Vertically Partitioned Data :Here the problem of Association Rule Mining where the transaction are distributed across sources were discussed. The two-party algorithm used for efficiently discovering frequent itemsets with minimum support levels, without revealing individual transaction values from either site. By vertically partitioned, each site contains some elements of a transaction. Using the traditional “market basket” example, one site may contain clothing purchases , while another has grocery purchases. Using a key such as credit card number and date, join these to identify relationships between

purchases of clothing and groceries. However, this discloses the individual purchases at each site, possibly violating consumer privacy agreements. Other problem is to mine association rules across two databases, where the columns in the table are different sites, splitting each row. One databases is designated the primary and it is the initiator of the protocol. The other databases is the responder. There is a join key present in both databases. The other remaining attributes are present in any onedatabases, but not both. The goal is to find association rule involving attributes other than the join key. Finally, it is necessary to quantify the accuracy and the efficiency of the algorithm, in view of the security restrictions.

Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data :The overview of Privacy Association Rule Mining has been outlined. The two phases are discovering candidate itemsets and determining which of the candidate itemsets meet the global support/confidence thresholds. The first phase uses commutative encryption where each party carries out encryption for its own frequent itemsets. The encrypted itemsets are then passed to other parties, until all itemsets are encrypted. These are passed to a common party to avoid and eliminate duplicates, and to begin decryption process. These set are then passed to each party, and each party decrypts each itemset. By this cryptographic technique for preventing the leakage of information have been studied [2].

The rest of the paper is organized as follows: In section III the design of the association rule with subgroup discovery is presented and provides the detailed description of modules used in the system. Section IV covers the conclusion part of this paper.

III SYSTEM DESIGN

The system focuses to build an association rule with cryptographic technique as well as rapid mining through the subgroup discovery technique. The modules of the system design include:

- Collecting the everyday transaction of the customer from various super markets and forming homogenous databases for analysis.
- Building a strong association rule with two significance measure namely minimum support and minimum confidence.
- Cryptographic technique is applied for datasets to perform the mining process in secure manner.
- Subgroup discovery technique is utilized for faster mining which brings top k patterns.

The input datasets are collected from marketing field by the point of sale to generate association rule through apriori algorithm. In this algorithm, the candidate key is generated for pruning the unwanted dataset before scanning the entire databases.

The following figure illustrates the overall work flow of the system:

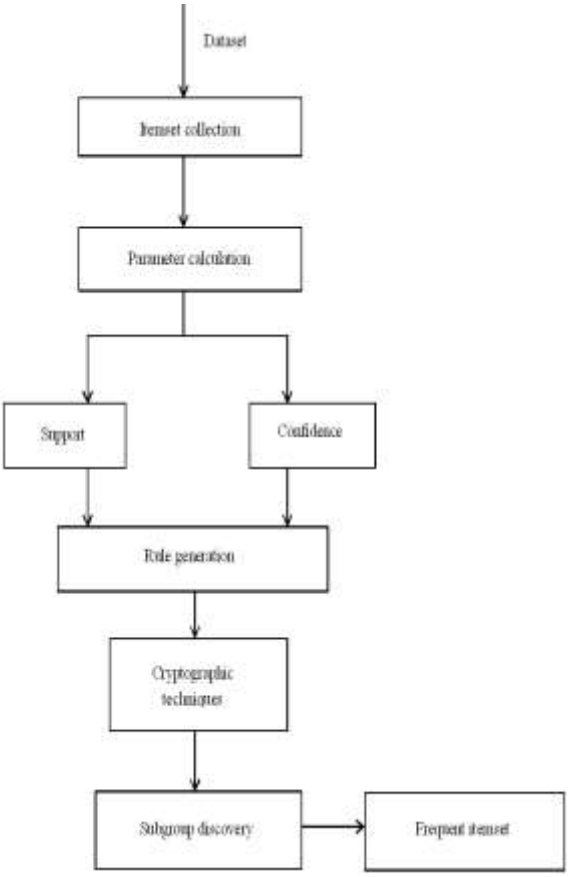

Fig. 1: Architecture Diagram

3.1 Collection of Datasets

Here the datasets refer to the everyday purchase transaction of customer. The system build the homogenous databases (same schema with unique entities) to feed the observed transaction data from the point of sale in various market.

3.2 Association Rule Mining

Association rules are if/then statements that help uncoverrelationships between seemingly unrelated data in a relational database or any other information repository. Anexample association rule "If a customer buysa dozen eggs, he is 80% likely to also purchase milk."By analyzing the data association rules are created forfrequent if/then patterns and using the criteria support andconfidence to identify the most important relationships. Support is an indicating term, it indicates how frequently the items appearin the database. Confidence is an indication of the number of timesthe if/then statements have been found to be true. Some technique used in building the association rule namely Apriori and Fast Distributed Mining Algorithm.

Apriori is designed to operate on databases containing transactions. The Apriori Algorithm is used to findassociations between different sets of data. It is sometimesreferred to as "Market Basket Analysis". Each set of data has a number of items called a transaction. The Apriori algorithm outputs sets of rules that tell us how often items arecontained in sets of data.

The FDM algorithm proceeds as follows:

1. Initialization
2. Candidate Sets Generation
3. Local Pruning
4. Unifying the candidate item sets
5. Computing local supports
6. Broadcast Mining Results

The two parameters for association rule has been calculated by following formula :

$$\text{Confidence} = \text{Support\_count}(A \cup B) \; \text{Support\_count}(A)$$

3.3 Advanced Encryption Standard

AES is not exactly Rijndael where Rijndael supports a larger range of block and keysizes.AES has a fixed block size and has a size of 128-bits and a key size of 128,192,256 bits.AES is a specification for the encryption of electronic data established by the U.S National Institute of Standard and Technology(NIST) in 2001.

For analyzing the structure and design of new AES ,the following three criteria were used:

- Resistance against all known attacks.
- Speed as well as code compactness on a wide range of platform.
- Design simplicity along with its similarities and dissimilarities and other symmetric ciphers.

AES algorithm is a symmetric block cipher algorithm that can encrypt(encipher)and decrypt(decipher)the information.AES can encrypt data much faster than Triple DES.AES is included in the ISO/IEC 18033-3 standard.AES ismade available in many different encryption packages,and is one of the first publicly accessible and open cipher approved by National Security Agency(NSA) for top secret information.

AES uses 128 bits blocks and key size of 128,192 or 256 bits. It doesn't have fiesel structure and it is block cipher which consists of 10 rounds with four separate functions namely byte substitution, permutation, arithmetic operation and Xor with a key.

The exact transformations occur as follows: each round consists of four steps:

- Add subkey: A portion of a key unique to this round is XOR with the round result. This operation provides confusion and incorporates the key.
- Byte Substitution: It uses S-box structure similar to DES, substituting each byte of a 128-bit block.
- Shift row: It is a simple permutation operation. For 128 and 192 bit block sizes, row n is shifted left circular (n – 1) bytes while for 256-bit blocks, row 2 is shifted 1 byte and row 3 and 4 are shifted 3 and 4 bytes, respectively.
- Mix column: The four bytes of every column are mixed in a linear fashion. This involves shifting left and XOR with the round result. These provide both confusion and diffusion.

The following figure illustrates the round involved in the AES algorithm:

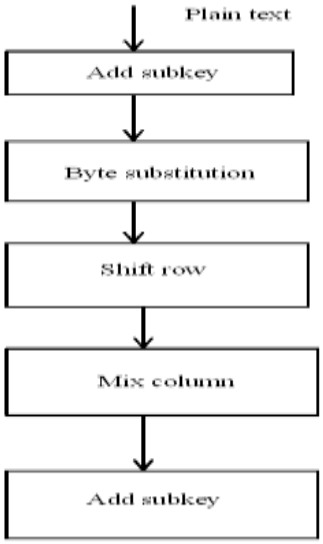

Fig. 2: Round 1 of AES Algorithm

AES has similar operation for both encryption and decryption where the operation get reserved while implementing decryption.

3.4 Subgroup Discovery

A subgroup discovery task mainly relies on the following fourproperties: the target variable, the quality function, the subgroup description language and the search strategy. The target variable (e.g., coronary heart disease) may be binary, nominal or numeric. Depending on its type, there are different analytic questions, e.g., we can search for significant deviations of the mean of a numeric target variable. The description language specifies the individuals from the general population belonging to the subgroup. Subgroup descriptionlanguages can be either single-relational or multi-relational. In the case of single-relational propositional languages a subgroup description can be defined as follows: Let $\Omega_A$the set of all attributes with an associated domaindom(a) of values.V is defined as the (universal) set of attribute values of the form (a = v), $a \in \Omega_A$, $v \in$dom(a) [6].

A quality function measures the interestingness of the subgroup mainly based on a statistical evaluation function, such as the the chi-squared statistical test. It is used by the search method to rank the discovered subgroups during search. Then, quality functions can be used to measure the characteristics of the subgroups according to the analytical questions. In the simplest case, one population share is considered, but also several shares (segments) can be used, e.g., segmenting by sex = male vs. sex = female.

Algorithm 1:

INPUT: length limit L and local databases D1,D2.....DS.
Site 1 initiates the secure calculation of |D| and $|D^+|$ and broadcasts the result
site1 creates a local iterator iter and queue Q_1, site S creates a local queue Q_S
whilehasnext(iter)do
site1 calculate and broadcasts S_i=next(iter)
site1 generates a random number r_i uniformly in [0,....M], enqueuesr_i in Q_1,adds its local support (|D1+[Si]|(1-P0)-|D1-[Si]|P0˙|D| to $r_i$and sends the result(mod M)to site2

sites 2.....s-1 add their local support to the intermediate sum and send the result (mod M)to the next site
site S adds its local support to the sum and enqueues the result,$q_i + r_i$(mod M),in Qs
end while
while Q1 contains more than 1 value *do*
site 1 dequeuesr_α and r_β from Q_1,generates a random number r' uniformly in [0,....M]and enqueues r' in Q_1
site1 generates and encrypts a circuit that computes(max(q_α,q_β)+r')modM) from.It sends the circuit to siteS together with the cryptographic keys corresponding to the input bits for r_α,r_β and r'
site1 sends the encoding table for the remaining inputs to site T
site S dequeues(r_α+$q_{\alpha'}$) and (r_β+$q_{\beta'}$)from Qs, asks site T for the corresponding cryptographic keys, evaluates the encrypted circuit and enqueues the result,(r'+max(q_α,q_β)modM)in Qs
end while
site 1 and S calculate $q_{max}$by exchanging the two remaining values
for every subgroup descriptions Si do
if $q_i'+r_i \geq r_i+_{qmax}\cdot|D|$ then return ‹si,qmax›
end for

IV CONCLUSION

Thus the subgroup discovery improves the performance of Association rule in horizontally distributed databases which brings top k pattern which leads to rapid mining. It also eliminate the duplicate item set. AES helps in preserving the confidential information of the organization. By this association rule mining can be built more strongly.

REFERENCES

[1] TamirTassa,"Secure Mining of AssociationRules in Horizontally Distributed Databases," IEEE trans. Knowledge and Data Eng., vol.26, no.4, pp.970-98, April 2014.

[2] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037, Sept. 2004.

[3] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, "A Fast Distributed Algorithm for Mining Association Rules," Proc. Fourth Int'l Conf. Parallel and Distributed Information Systems (PDIS), pp. 31-42, 1996.

[4] J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge discovery and Data Mining (KDD), pp. 639-644, 2002.

[5] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," Proc. Crypto, pp. 175-186, 2005.

[6] H. Grosskreutz, B. Lemmen, and S. Ruping, "Secure Distributed Subgroup Discovery in Horizontally Partitioned Data," Trans. Data Privacy, vol. 4,no. 3,pp. 147-165,2011.